Published in "Applied Ergonomics 40 : 670-677, 2009
which should be cited to refer to this work.

The influence of prototype fidelity and aesthetics of design in usability tests: Effects on user behaviour, subjective evaluation and emotion

Juergen Sauer*, Andreas Sonderegger

Department of Psychology, University of Fribourg, Rue de Faucigny 2, 1700 Fribourg, Switzerland

An empirical study examined the impact of prototype fidelity on user behaviour, subjective user evaluation and emotion. The independent factors of prototype fidelity (paper prototype, computer prototype, fully operational appliance) and aesthetics of design (high vs. moderate) were varied in a between-subjects design. The 60 participants of the experiment were asked to complete two typical tasks of mobile phone usage: sending a text message and suppressing a phone number. Both performance data and a number of subjective measures were recorded. The results suggested that task completion time may be overestimated when a computer prototype is being used. Furthermore, users appeared to compensate for deficiencies in aesthetic design by overrating the aesthetic qualities of reduced fidelity prototypes. Finally, user emotions were more positively affected by the operation of the more attractive mobile phone than by the less appealing one.

1. Introduction

1.1. Prototype fidelity

Product designers are typically faced with the problem that human behaviour in operating a system needs to be predicted although the system has not yet been fully developed. The system may only be available in a rudimentary form, which falls well short of a fully operational prototype. This may range from specifications (descriptions based on requirement analysis) through cardboard mock-ups to virtual prototypes.

The question of which prototype is to be used for usability testing is strongly influenced by a number of constraints that are present in industrial design processes, notably time pressure and budgetary limitations. This usually calls for the use of low fidelity prototypes (e.g., paper prototype) because they are cheaper and faster to build. Although prototypes of various forms are widely used in industry, there is little comparative research on the utility of prototypes at different fidelity levels. A review of the research literature has revealed a total of nine studies in which comparative evaluations of different prototypes were carried out (Sefelin et al., 2003; Virzi et al., 1996; Säde et al., 1998; Nielsen, 1990; Catani and Biers, 1998; Walker et al., 2002; Wiklund et al., 1992; Hall, 1999; Sauer et al., 2008). The majority of studies concluded that the reduced fidelity prototypes provided equivalent results to fully operational products. Only three studies (Nielsen, 1990; Hall, 1999; Sauer et al., 2008) reported some benefits of higher fidelity prototypes over lower fidelity prototypes.

The decision of selecting a prototype for human factors testing entails a dilemma. On the one hand, a prototype of too high fidelity is very time-consuming and expensive to build; hence valuable resources are wasted. On the other hand, the findings obtained with a prototype of too low fidelity may not be valid. This requires the careful consideration of what level of fidelity would be best to opt for. The concept of prototype fidelity is quite broad in scope, encompassing a number of different dimensions upon which a prototype can differ from the reference product. Virzi et al. (1996) have suggested a classification system that distinguishes between four dimensions of fidelity: degree of functionality, similarity of interaction, breadth of features, and aesthetic refinement.

Degree of functionality is concerned with the level of detail to which a particular function has been modelled. For example, the user–product dialogue for taking a picture with a mobile phone can be modelled in its entirety or in a reduced form. *Interactivity* refers to the type of interface (i.e. controls and displays) with which the prototype is modelled. For example, on a computer-based simulation of a telephone, one may use a touch screen to enter a phone number directly with the fingers (higher fidelity) or use a mouse to do the same on a conventional screen (lower fidelity). *Breadth of functions* refers to the extent to which all functions of the target product are modelled in the prototype (e.g., 4 out of 5 displays and 3 out of 4 control elements of the real system are represented in the prototype). *Aesthetic refinement* refers to the extent to which there are similarities between the prototype and the target product with

* Corresponding author. Tel.: +41 26 3007622; fax: +41 26 3009712.
E-mail address: juergen.sauer@unifr.ch (J. Sauer).

regard to physical properties, such as shape, size, colour, texture and material. This dimension has also been referred to as the 'look' of the prototype (e.g., Snyder, 2003). The model of Virzi et al. (1996) clearly indicates that a prototype can differ from the reference product in many different aspects. Overall, the model of Virzi et al. may represent a useful framework for designers to guide the prototype development process.

1.2. Usability testing

In order to assess the utility of prototypes, usability tests are often used since they allow for user–product interaction to be measured under controlled conditions. The ISO Standard of usability (ISO 9241-11) refers to the three main aspects of usability: effectiveness, efficiency, and user satisfaction. Effectiveness and efficiency may be considered objective measures since they examine actual user behaviour while user satisfaction refers to subjective measures that take into account the user's opinion and feelings.

1.2.1. User behaviour

Effectiveness refers to the extent to which a task goal is successfully achieved with the product (Jordan, 1998a). This may be measured by the proportion of users that can actually complete a given task. In addition to rate of task completion, effectiveness may also be measured by the quality of the output (e.g., taste of a cup of coffee brewed with a coffee maker). *Efficiency* refers to the amount of resources expended to accomplish a task goal (Jordan, 1998a). Typical measures of efficiency are deviations from the critical path (e.g., number of superfluous clicks on a menu during task completion), error rates (e.g., number of wrong commands), and time on task (e.g., time needed to accomplish the task).

All these measures may be taken during usability tests. However, knowledge about the influence of different levels of prototype fidelity on these outcome measures is limited. Most of the studies cited in the literature review above focused on usability problems alone, with a smaller number of studies also measuring user satisfaction (e.g., Catani and Biers, 1998; Wiklund et al., 1992). The review of the studies also suggests that empirical research has concentrated very much on effectiveness measures, with efficiency issues being somewhat neglected. The focus on usability errors may have contributed to a largely positive evaluation of prototypes of lower fidelity in usability tests, which might not be entirely justified. It remains to be empirically tested whether this positive evaluation can still be maintained when a wider range of measures of user behaviour is examined.

1.2.2. Subjective user evaluations and emotions

In addition to objective data, data on user satisfaction are often collected during usability tests by means of standardised questionnaires and semi-structured interviews. The questionnaires range from rather short instruments (e.g., 10-item Software Usability Scale of Brooke, 1996) to very elaborate instruments that measure different facets of user satisfaction (e.g., Questionnaire for User Interaction Satisfaction containing 71 questions; Chin et al., 1988). These questionnaires have been typically employed on fully operational products so that it remains to be seen to what degree reduced fidelity prototypes provide valid data to estimate user satisfaction with the real product.

While user satisfaction has been a notion in usability testing for some time, more recently consumer product design has also become concerned with concepts such as joy, pleasure and fun (Norman, 2004b; Jordan, 1998b, 2000). While the concept of satisfaction may be considered an attitude towards the product (i.e. like the concept of job satisfaction in a work context; e.g., Schleicher et al., 2004), joy, pleasure and fun (which appear to be used largely synonymously in the usability literature) represent emotions, which, in contrast, have a clear focus on the internal state of the user. Emotions are increasingly considered to be an important issue in consumer product design, as a rising number of publications have paid testimony to (e.g., Helander and Khalid, 2006; Norman, 2004a; Brave and Nass, 2003). For example, there is evidence that the emotional response to a product is more influential than cognitive components in determining consumer decision-making (Shiv and Fedorikhin, 1999). Emotions are also of particular interest because they represent a faster and more immediate reaction to an object than a pure cognitive evaluation (Khalid, 2006).

Concerning the effects of prototype fidelity, it is of particular interest to what extent emotions associated with product utilisation can be predicted from low and medium fidelity prototypes. In order to assess the user's emotional response, product developers typically use prototypes of higher fidelity for this purpose (e.g., 3D mock-up), which are characterised by considerable aesthetic refinement. This is due to concerns that lower fidelity prototypes (e.g., involving only a rough sketch of the design) would not elicit the same emotional response. If a prediction of the emotional response was possible on the basis of a prototype with reduced fidelity, it would allow designers to measure the impact of a product on user emotions at an earlier stage in the design process rather than having to wait until an aesthetically refined prototype can be made available.

Closely related to emotions is the aesthetic appeal of a product. There are a number of concepts in the research literature that refer to the exterior properties of a product and the user's response to it, such as aesthetics, appearance, attractiveness and beauty (e.g., Hekkert et al., 2003; Chang et al., 2007; Hassenzahl, 2004). However, these concepts are not employed consistently across research communities and research fields. For example, with regard to the concept of aesthetics, Lavie and Tractinsky (2004) have distinguished between the factors classical and expressive aesthetics while Hekkert et al. (2003) have identified novelty and typicality as factors. Other work considers the term aesthetics as the user's response to the appearance of the product (Crilly et al., 2004). In the present article, we will use the term aesthetic of design to refer to the visual appearance of a product (i.e. independent variable) whereas the users' response to these product properties is referred to as attractiveness (i.e. dependent variable).

Aesthetics of product design has long been considered an important issue in the field of industrial design (e.g., Yamamoto and Lambert, 1994). However, in the field of ergonomics, only more recently there have been calls for a stronger consideration of aesthetics as a pertinent factor in system design in addition of safety, usability and comfort (e.g., Liu, 2003). While aesthetics has also been linked to consumer decision-making, its influence may not be limited to that field since it may also affect the perceived usability of products. For example, research has indicated that aesthetic products are perceived as being more usable than less appealing ones (Tractinsky, 1997). This finding suggests that the influence of aesthetics is not limited to the product's appeal to the user but also affects usability ratings and, possibly, the way the product is being used.

1.3. The present study

The review of the literature revealed that there is only little work that examined the effects of prototypes' fidelity on efficiency measures, user satisfaction, emotions and attractiveness. The limited work available mainly focussed on effectiveness measures (e.g., number of users that were able to complete the task). Against this background, the main research question examines the extent to which data obtained in usability tests with prototypes of reduced fidelity allow the prediction of user responses (i.e. observed

behaviour and subjective evaluations) to the fully operational system. This was investigated by comparing paper and computer-simulated prototypes with fully operational products. A subsidiary research question was concerned with the aesthetic appeal of the design and to what extent it may modify the relationship between prototype fidelity and user responses.

The mobile phone was used as a model product. This appliance was regarded as particularly suitable for the purpose of this study because it is not only functionality and usability that are important for this product group. A mobile phone may be considered a lifestyle product to which a certain prestige value is attached, which may trigger off stronger emotional reactions during user–product interaction than a conventional product. The measures taken in this study covered the main outcome variables of a usability test. This included various performance measures as well as subjective measures ranging from usability ratings to emotional states.

Based on the research reviewed, the following research assumptions were formulated:

(a) User performance would be higher for the fully operational product than the two reduced fidelity prototypes (task completion time and efficiency of operation).
(b) The difference in user behaviour and subjective usability ratings between the fully operational product and reduced fidelity prototypes would be larger for the paper prototype than for the computer-based prototype since the latter is more similar to the fully operational product.
(c) An aesthetically more appealing appliance would create more positive emotions and would receive higher usability ratings than a less appealing product.
(d) For the fully operational product, the effects of design aesthetics on emotion and subjective usability would be more pronounced than for the reduced fidelity prototypes (i.e. interaction *fidelity level* × *appliance usability*). This is because a less appealing aesthetic design would be more tolerable to users on an unfinished prototype than on a fully operational product with a finalised design.

2. Method

2.1. Participants

Sixty participants (58.3% male, 41.7% female) took part in the study, aged between 19 and 41 yrs (M = 23.8 yrs). They were students of the University of Fribourg and all of them were regular users of a mobile phone. A strict selection criterion was that participants should not have been familiar with the particular mobile phone they were going to use in the study. Participants were not paid for their participation.

Some of the participants had, however, experience with other models of the same brand they used in the experiment. In total, 23 participants were found to have such previous experience. However, post-hoc tests comparing participants with and without previous brand experience showed no difference for any of the dependent variables (all $t < 1$), suggesting no significant influence of this factor.

2.2. Experimental design

A 3 × 2 between-subjects design was employed in the study. The main independent variable *prototype fidelity* was varied at three levels: paper prototype, computer-based prototype, and fully operational appliance. A second independent variable *aesthetics of design* was manipulated at two levels: highly appealing vs. moderately appealing (see Section 2.4.1). Each participant was randomly assigned to one of the six experimental conditions.

2.3. Measures and instruments

2.3.1. User behaviour

Two measures of user behaviour were recorded: *Task completion time* (s) referred to the time needed to accomplish the task. *Interaction efficiency* was a composite parameter, dividing the optimal number of user inputs by the actual number of user inputs.

2.3.2. Subjective usability evaluation

The German-language questionnaire *Multimetrix*^{S} (Ollermann, 2001) was employed to measure usability ratings of the user. This instrument was largely based on the design principles suggested by the ISO Standard (ISO 9241-11). The questionnaire was slightly modified by removing items that were irrelevant for the intended application area (e.g., the subscales "media quality" and "suitability of individualisation" were removed since they were considered not to be applicable). This reduced the number of items from 86 to 58. The statements had to be rated on a 5-point Likert scale (agree, partly agree, neither agree nor disagree, partly disagree, disagree). If the item was not applicable, the user was given the choice to tick the appropriate category. The psychometric properties of the Multimetrix are sufficient, with Cronbach's alpha ranging from .63 to .89 for the different scales (Willumeit et al., 1995). The subscales of the instrument were as follows:

- Suitability for the task (example item translated from German: "The system forces me to carry out unnecessary actions").
- Conformity with user expectations ("Messages of the software always appear at the same place").
- Information and information structure ("The software contains all relevant information").
- Suitability for learning ("The functions of the software can be easily learnt").
- Self-descriptiveness ("I can use the software straight away without the help from others").
- Controllability ("I feel that I have control over the software at any time").
- Error tolerance ("Correcting errors involves little effort").
- User acceptance ("The software is overloaded with graphical design features").

2.3.3. Emotions and attractiveness

2.3.3.1. Learning affect monitor (LAM). This is a 32-item questionnaire developed by Reicherts et al. (2005) to capture emotions experienced in daily life. It was slightly adapted to make it suitable for the purpose of the present study. Only a subset of 10 items were employed and analysed, excluding those items that were considered to be less relevant for user–product interaction. The items had a 9-point Likert scale ranging from "not at all" to "very much". The selection of items was based on the emotions covered by PrEmo (Desmet, 2003). PrEmo is an instrument that aims to measure emotions relevant to consumer product evaluation by using a cartoon character that depicts each emotion during a short animation. The selected items were identical or very similar to the set of 14 emotions measured by PrEmo (excluding four emotions for which no equivalent emotion had been found in the LAM instrument). The remaining 10 items referred to the following emotions: irritation, boredom, disappointment, delight, enthusiasm, surprise, contentment, disgust, anger, and happiness.

2.3.3.2. Attractiveness. The attractiveness of the product was measured on a one-item 5-point Likert scale, with the item being phrased: "The design of the mobile phone is appealing" (agree, partly agree, neither agree nor disagree, partly disagree, disagree). The item, translated from German, was self-developed and

intended to capture very broadly the user's response to the aesthetic design of the product.

2.4. Materials

2.4.1. Mobile phones (high fidelity prototype)

Two mobile phones (SE W800i from Sony Ericsson and M V3690 from Motorola) were selected for the study (see Fig. 1a). The SE W800i (launched onto the market in the year 2005) was considered to be an aesthetically appealing appliance whereas the M V3690 (launched in 1999) was chosen as a model for a moderately appealing appliance. The selection of the two mobile phones was based on expert judgement, involving the two authors and two other raters who independently rated a total of 15 telephones for aesthetic appeal. The two appliances with the most extreme ratings at either end were selected for the experiment. The manipulation check in the experiment was successful, since it was later confirmed by the participants who rated the two telephones very differently for their attractiveness (see Section 3.3).

2.4.2. Touch screen computer (medium fidelity prototype)

For the medium fidelity condition, computer-based simulations of the dialogue structure of each mobile phone were developed by using Microsoft PowerPoint (see Fig. 1b). Each prototype allowed the user to interact with the mobile phone and to carry out the same tasks as with the real product. For the purpose of the study, only the top two levels of the dialogue structure were fully developed for all functions rather than providing an emulation of the complete functionality of each mobile phone. The dialogue structure was only modelled in full depth for the task-relevant menu items. If the user left the optimal dialogue path by more than two levels of the menu structure, an error message was displayed ("Wrong path, please go back"). To obtain the sketchy appearance often found for prototypes employed in usability tests, both icons and text were drawn by hand on a graphic tablet using an electronic pen. The simulation was run on an IBM ThinkPad ×41 Tablet PC with a touch screen, which enabled the user to interact directly with the prototype instead of having to use a mouse. This ensured that a similar kind of interface (i.e. high interactivity) is used for the prototype compared to the real product (cf. McCurdy et al., 2006).

2.4.3. Paper prototype (low fidelity prototype)

The paper prototype consisted of a collection of cards (sized 90 mm × 180 mm for Sony Ericsson and 80 mm × 210 mm for Motorola) upon which all configurations were printed that were modelled by the computer simulation (see Fig. 1c). These were basically exact replications of the different screen shots. The cards were kept in an indexed card box and presented to the user by the experimenter during the usability test. The user performed the task by pointing the finger to one of the buttons on the paper prototype. Based on the user's selection, the experimenter presented the card reflecting the change in display content initiated by the action.

2.5. User tasks

For the usability test, two user tasks were chosen. The first task ("text message") was to send an already prepared text message to another phone user. This represents a task frequently carried out by a typical user. The second task ("phone number suppression") was to suppress one's own phone number when making a call. This task is a low-frequency task compared to the first and was therefore considered to be slightly more difficult.

The tasks differed slightly with regard to the number of commands entered by the user to complete the task successfully. For the "text message" task, this was 8 inputs for the mobile phone from Sony Ericsson and 13 inputs for the Motorola phone. For the "phone number suppression" task, the optimal way of completing the task consisted of 14 inputs (Sony Ericsson) and 8 inputs (Motorola). The tasks were always presented in the same order, beginning with the "text message" task and followed by the "phone number suppression" task.

2.6. Procedure

The study was conducted in a usability laboratory at the University of Fribourg. After welcoming the participant and providing a briefing about the purpose of the experiment, a biographical questionnaire was administered, followed by the LAM questionnaire to obtain a baseline measure of the participant's emotional state. Participants were randomly assigned to one of the experimental conditions (if a participant had already gained some experience with that particular mobile phone, the participant was removed from that experimental condition). The next activity of participants was the completion of the two experimental tasks (see Section 2.5). During the entire testing procedure, an experimenter was present and took notes. Immediately after the two tasks had been completed, the emotions of the participant were measured again with the LAM questionnaire. This was followed by the presentation of the one-item attractiveness scale and the Multimetrix questionnaire. Finally, the participant was given the opportunity to provide feedback to the experimenter about the prototype and the testing procedure.

3. Results

3.1. User behaviour

3.1.1. Task completion time

This measure was not taken for the paper prototype since it would not have been an adequate reflection of user performance. The measure would have been confounded with the response time

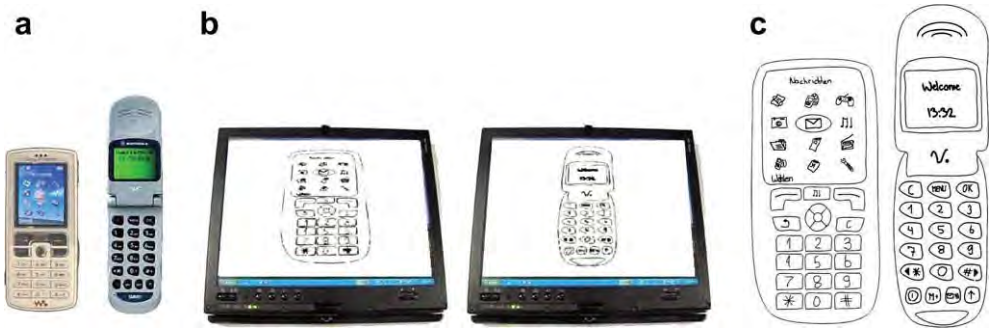

Fig. 1. Prototypes of mobile phone: (a) high fidelity, (b) medium fidelity, (c) low fidelity.

Table 1
Measures of user behaviour as a function of prototype fidelity and aesthetic design of appliance (N/A: not available).

	Paper prototype (low fidelity)	Computer-based prototype (medium fidelity)	Fully operational appliance (high fidelity)	Overall
Task completion time (s)	N/A	268.8	157.2	
Highly aesthetic design	N/A	342.1	138.4	240.2
Moderately aesthetic design	N/A	195.6	175.9	185.7
Interaction efficiency index	.58	N/A	.58	
Highly aesthetic design	.61	N/A	.63	.62
Moderately aesthetic design	.55	N/A	.53	.54

of the human "playing" the computer. Overall, the data showed a strong between-participant variation with regard to this performance measure (e.g., task completion times ranged from 46 to 498 s). The analysis revealed a main effect of prototype fidelity (see Table 1), with users requiring significantly more time when using the computer-based prototype than the fully operational appliance ($F = 9.72$; df = 1,36; $p < .005$). This main effect was modulated by a significant cross-over interaction between fidelity and appliance aesthetics ($F = 6.59$; df = 1,36; $p < .05$). While in the condition "computer prototype/highly aesthetic design", the completion time was longest, the same model showed the fastest task completion times when the fully operational appliance was used ($F = 6.59$; df = 1,36; $p < .05$; LSD-tests: $p < .05$). No significant main effect of appliance aesthetics was found ($F = 2.31$; df = 1,36; $p > .05$).

3.1.2. Interaction efficiency

The results of the efficiency of user–product interaction (i.e. optimal number of user commands divided by actual number of user commands) are presented in Table 1. Due to a failure of the data logging facility of the computer prototype, the number of user–system interactions was not accurately counted so that no data were available for this experimental condition. For the remaining conditions, no differences between cells were found. This was confirmed by a two-factorial ANOVA, which showed no effect of prototype fidelity ($F < 1$), none of design aesthetics ($F = 2.51$; df = 1,36; $p > .05$), and no interaction between the two factors ($F < 1$). In the medium fidelity condition, the experimenter made the interesting observation that many users clicked several times directly on the display of the mobile phone presented on the computer touch screen rather than the buttons until users realised that only the computer had a touch screen but not the simulated mobile phone. This type of error related to prototype interactivity was not observed under the paper prototype condition.

3.2. Subjective usability evaluation

A multivariate analysis of variance (MANOVA) was carried out to test for overall effects of the independent variables on 8 rating scales of the Multimetrix. The MANOVA showed an overall effect for appliance aesthetics ($F = 12.4$; df = 7,48; $p < .001$) but not for prototype fidelity ($F < 1$) and no interaction was found ($F = 1.05$; df = 14,96; $p > .05$). Separate analyses on each scale revealed that the highly appealing design (i.e. SE W800i) was given higher usability ratings than the moderately appealing design on all 8 scales (all scales were strongly correlated with each other, suggesting that users did not distinguish much between them) as well as on the overall scale. All effects were highly significant, as the data in Table 2 demonstrate. The absence of an effect of prototype fidelity suggests that fidelity does not influence the perceived usability of a product.

3.3. Emotions and attractiveness

3.3.1. Emotions

A MANOVA was carried out on the 10 LAM items. The analysis revealed no effect for fidelity level ($F < 1$) but an effect for design aesthetics was observed ($F = 2.86$; df = 10,45; $p < .01$). No interaction between the two factors was recorded ($F < 1$). In Table 3 the means of participant ratings at t_0 (i.e. prior to usability test) and t_1 (i.e. after usability test) are presented as a function of design aesthetics. Separate univariate analysis of variance on single items revealed significant effects for five items. The strongest effect was found for 'delighted', followed by 'disappointed', 'happy', 'irritated', and 'angry'. No significant differences were found for the five other emotions. The data in Table 3 also indicated that the emotion "surprised" showed a very strong increase from t_0 to t_1 for both appliances ($F = 16.1$; df = 1,59; $p < .001$).

3.3.2. Attractiveness

The ratings of the attractiveness scale are presented in Table 4. As expected, the analysis revealed a strong main effect of aesthetics, with the highly aesthetic appliance being given higher ratings ($F = 25.3$; df = 1,53; $p < .001$). This demonstrated that the experimental manipulation had been successful. More interesting was the interaction between prototype fidelity and design aesthetics ($F = 4.6$; df = 2,53; $p < .05$), with the moderately aesthetic design of the fully operational appliance having a significantly lower rating than all the other conditions (LSD-test: $p < .005$). No significant difference was found between the two paper prototypes and the two computer prototypes (both LSD-tests: $p > .05$). Finally, a main effect of prototype fidelity was found ($F = 3.4$; df = 2,53; $p < .05$),

Table 2
User ratings on overall scale and each subscale (1–5) of Multimetrix^S (***$p < .001$).

	Highly aesthetic design	Moderately aesthetic design	Results of analysis of variance [F (1,54)]
Overall scale	3.81	2.88	77.3***
Suitability for the task	3.75	2.77	46.3***
Conformity with user expectations	4.00	3.41	21.7***
Information and information structure	3.73	2.37	73.8***
Suitability for learning	3.92	2.87	27.6***
Self-descriptiveness	3.71	2.37	80.7***
Controllability	4.14	3.15	33.2***
Error tolerance	3.23	2.74	14.7***
Acceptance	4.04	3.36	22.9***

Table 3
Mean ratings of emotions at t_0 (prior to usability test) and t_1 (after usability test) as a function of appliance usability on a 9-point Likert scale; significant differences as a function of design aesthetics are indicated by stars ($^{*}p < .05$, $^{***}p < .001$).

	Highly aesthetic design			Moderately aesthetic design		
	t_0 (SD)	t_1 (SD)	Difference ($t_1 - t_0$)	t_0 (SD)	t_1 (SD)	Difference ($t_1 - t_0$)
Irritated*	2.80 (1.9)	2.33 (1.4)	−.47	2.53 (1.5)	3.27 (2.0)	+.73
Bored	2.37 (1.7)	2.3 (1.6)	−.07	3.0 (1.8)	2.56 (1.5)	−.44
Disappointed*	2.33 (2.0)	1.9 (1.3)	−.43	1.70 (1.1)	2.3 (1.7)	+.83
Delighted***	5.80 (1.5)	6.43 (1.6)	+.63	5.73 (1.6)	5.1 (1.7)	−.63
Enthusiastic	5.2 (2.0)	5.0 (2.2)	−.20	4.57 (1.8)	4.27 (1.8)	−.30
Surprised	2.53 (1.5)	3.47 (2.1)	+.94	2.27 (1.5)	3.46 (2.1)	+1.19
Contented	6.3 (1.9)	5.93 (2.1)	−.37	6.07 (1.6)	5.4 (1.8)	−.67
Disgusted	1.57 (1.5)	1.43 (1.2)	−.14	1.27 (.8)	1.3 (.7)	+.03
Angry*	2.0 (1.7)	1.63 (1.4)	−.36	1.86 (1.2)	2.1 (1.6)	+.23
Happy *	5.63 (2.1)	6.13 (2.1)	+.50	5.76 (1.7)	5.23 (1.7)	−.53

which was only due to the low score of the real appliance with the moderately aesthetic design.

4. Discussion

The central question of this article concerned the utility of prototypes that are of lesser fidelity than the reference system during usability tests. The main results showed that task completion time may be overestimated when a computer-based simulation is used. Furthermore, the effects of fidelity levels on attractiveness ratings appeared to be stronger for less appealing products than for attractive ones. It also emerged that objective performance parameters collected during the usability test and subjective usability ratings were not associated. Finally, the results showed no evidence for fidelity level affecting emotions or subjective user evaluation.

The results showed that task completion time was higher for the computer-based simulation than when a fully operational product was being used. This effect was observed for both mobile phones, though they differed with regard to the strength with which this effect occurred. The increased task completion time under the computer condition was partly caused by prototype-specific errors being made by users that resulted from differences in the interactivity of prototype (cf. model of Virzi et al., 1996). With the computer prototype, ineffective clicks were made by participants because they erroneously extended the interactivity of the device from the computer screen (direct manipulation was possible) to the display of the simulated mobile phone (direct manipulation was not possible). However, it was only the representation of the mobile phone's controls on the touch screen that was interactive. Although the touch screen permits a more natural interaction of the user with the mobile phone than a conventional screen (for which the user needs to use a mouse), this advantage may be accompanied, as observed in the present case, by unanticipated side effects in the form of negative transfer.

For the attractiveness rating of the appliances, an interesting interaction between prototype fidelity and design aesthetics was observed. While there was no difference in ratings across different fidelity levels for the highly aesthetic mobile phone, the moderately aesthetic phone was rated lower on attractiveness for the original appliance than for the reduced fidelity prototypes. The fact that the two reduced fidelity prototypes had similar ratings like the original appliance for the highly aesthetic design is in itself a somewhat surprising result. This suggests that some compensatory activity on the part of the user took place since neither the paper prototype nor the computer-based prototype was aesthetically refined (e.g., lacking colour and shape of the reference appliance). Users may have mentally anticipated of what the real appliance might look like and employed this mental picture as a basis for their rating. For the moderately aesthetic phone design, users may have engaged in a similar process in that they extrapolated the appearance of the computer and paper prototypes to the real appliance (indeed, there were no significant differences between the two computer-based prototypes and the two paper prototypes across phone types). Since the computer-based and paper prototypes were judged to be more attractive than the real appliance, it can be speculated that under the reduced fidelity conditions users created a mental model of the real appliance representing a much more attractive design than the real appliance actually enjoyed. This may suggest a kind of "deficiency compensation" effect. As this interaction between prototype fidelity and design aesthetics was not predicted, it needs to be treated with some caution but, if confirmed in subsequent studies, it would have implications for the use of reduced fidelity prototypes for the purpose of attractiveness judgements.

The results showed no association between objective performance parameters and subjective usability evaluation. While there was a clear preference of users for the more aesthetic appliance because of higher attractiveness ratings and higher perceived usability, this was not paralleled by better objective usability of that appliance. This suggests that perceived usability may be more strongly associated with attractiveness ratings than objectively measured usability parameters. This result is in support of the findings of Tractinsky (1997), who proposed that the beauty of design would positively affect perceived usability. While in Tractinsky's study no user–product interaction took place (with the usability rating of users being based on the mere look of the product), the present study provided similar evidence even for the case when user–product interaction occurred. If this finding was to be found consistently, it would imply that the beauty of a product was such an important aspect that it would also need to be considered by designers and engineering psychologist when designing for usability.

Table 4
User ratings of attractiveness of product (1–5) as a function of prototype fidelity and design aesthetics.

	Paper prototype (low fidelity)	Computer-based prototype (medium fidelity)	Fully operational appliance (high fidelity)	Overall
Overall	3.3	3.6	2.9	
Highly aesthetic design	3.7	4.0	4.0	3.9
Moderately aesthetic design	3.0	3.3	1.8	2.7

The changes in emotions during the usability test (i.e. from t_0 to t_1) were quite substantial, suggesting that user–product interaction constitutes a significant emotional experience. The intensity of the emotional experience may have been increased by two factors. First, the usability testing procedure that included the presence of an experimenter may have intensified the emotions recorded because of the increased arousal induced by the presence of others, as suggested by social facilitation theory (Cottrell et al., 1968). Second, it may be that at t_0 emotions were measured but at t_1 measurements of sentiments were taken. Sentiments refer to the user's feelings towards the appliance rather than reporting their internal state (Brave and Nass, 2003). These may have been evoked during product utilisation, resulting in a considerable change in user ratings. At t_0 users reported their general internal emotional state while at t_1 their self-reported state was closely linked to the directly preceding experience with product utilisation. This may explain the considerable changes across measurement points. Similar to the findings for attractiveness ratings, there was no evidence for a different emotional reaction being triggered off by reduced fidelity prototypes compared to the real appliance. The same was observed for subjective usability evaluation (i.e. even prototypes of lower fidelity seemed to be useful to assess subjective usability). Users may have achieved this by creating a mental model of the real appliance (under paper and computer prototype conditions) upon which their judgements are based.

The use of reduced fidelity prototypes raises the broader issue of validity of usability testing. Concerns have been expressed about the validity of usability tests, given the remarkable inconsistencies in test outcomes that were observed across tests (e.g., Lewis, 2006). While it is generally agreed that usability testing improves the usability of products (as opposed to not conducting any usability test), the validity of the test could be increased if we had a better understanding of the factors that influence validity. Of the many forms of validity, ecological and predictive validity may be of particular interest. In order to improve the ecological validity of a usability test (i.e. the extent to which behaviour in a test situation can be generalised to a natural setting), the influence of the wider testing environment needs to be considered (e.g., Brehmer and Dörner, 1993). This refers in particular to the physical and social aspects of test environment (e.g., lab set-up, presence of observers). For this purpose, a model (called the Four-Factor Framework of Contextual Fidelity) has been proposed, which explicitly refers to these factors (Sauer et al., submitted for publication). Predictive validity coefficients of paper and computer prototypes may also be determined in future studies, using a similar approach as in personnel selection where the validity of different selection methods has been determined. Test participants would first complete a set of tasks with a reduced fidelity prototype and subsequently (after a time interval) with a real product. Lastly, we would like to point out a methodological weakness of this study. This refers to the exhibition of the mobile phone's brand name in the high fidelity condition. The brand name was left uncovered to produce a more natural testing situation but it cannot be excluded that this may have influenced emotion and attractiveness ratings.

Finally, there is a need to carry out more research into the effects of prototype fidelity and design aesthetics to examine whether the findings of the present study can be replicated with modified design characteristics and also with different interactive consumer products. For example, it would be important to see whether the interaction found for attractiveness ratings can be replicated if the reduced fidelity prototypes had been aesthetically more refined instead of presenting a rough sketch. The question of which prototype should be used would not only be relevant in the context of usability testing but also when designers present prototypes of the work that was commissioned by their clients. In this situation, the issue of aesthetics is also of great importance since they may influence the client's decisions. Overall, the findings suggest that prototypes of reduced fidelity may be suitable for modelling the reference system. From the findings of the present work, it appears that in order to design a highly usable product, an appealing design would be one of the *necessary* product features. This would suggest that the issue of aesthetics should be closer to the heart of the ergonomic design process than perhaps previously thought.

Acknowledgements

We are very grateful to Amadeus Petrig for developing the computer prototype. Furthermore, we would like to thank two anonymous reviewers for their very helpful comments on an earlier version of this manuscript.

References

Brave, S., Nass, C., 2003. Emotion in human–computer interaction. In: Jacko, J., Sears, A. (Eds.), The Human–Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications. Lawrence Erlbaum Associates, Mahwah, pp. 81–96.

Brehmer, B., Dörner, D., 1993. Experiments with computer-simulated microworlds: escaping both the narrow straits of the laboratory and the deep blue sea of the field study. Computers in Human Behavior 9, 171–184.

Brooke, J., 1996. SUS: a 'quick and dirty' usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, I.L. (Eds.), Usability Evaluation in Industry. Taylor & Francis, London, pp. 189–194.

Catani, M.B., Biers, D.W., 1998. Usability evaluation and prototype fidelity: users and usability professionals. In: Proceedings of the Human Factors Society 42nd Annual Meeting, October 5–9, 1998, Chicago, USA. HFES, Santa Monica, pp. 1331–1335.

Chang, H.C., Lai, H.H., Chang, Y.M., 2007. A measurement scale for evaluating the attractiveness of a passenger car form aimed at young consumers. International Journal of Industrial Ergonomics 37 (1), 21–30.

Chin, J.P., Diehl, V.A., Kent, L.N., 1988. Development of an instrument measuring user satisfaction of the human–computer interface. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, May 15–19, 1988, Washington, USA. ACM, New York, pp. 213–218.

Cottrell, N.B., Wack, D.L., Sekerak, G.J., Rittle, R.M., 1968. Social facilitation of dominant responses by the presence of an audience and the mere presence of others. Journal of Personality and Social Psychology 9, 245–250.

Crilly, N., Moultrie, J., Clarkson, P.J., 2004. Seeing things: consumer response to the visual domain in product design. Design Studies 25 (6), 547–577.

Desmet, P.M.A., 2003. Measuring emotion: development and application of an instrument to measure emotional responses to products. In: Blythe, M.A., Overbeeke, K., Monk, A.F., Wright, P.C. (Eds.), Funology: from Usability to Enjoyment. Kluwer, Dordrecht, pp. 111–123.

Hall, R.R., 1999. Usability and product design: a case study. In: Jordan, P., Green, W.S. (Eds.), Human Factors in Product Design. Taylor & Francis, London, pp. 85–91.

Hassenzahl, M., 2004. The interplay of beauty, goodness, and usability in interactive products. Human–Computer Interaction 19 (4), 319–349.

Hekkert, P., Snelders, H.M.J.J., van Wieringen, P.C.W., 2003. Most advanced, yet acceptable: typicality and novelty as joint predictors of aesthetic preference. British Journal of Psychology 94, 111–124.

Helander, M.G., Khalid, H.M., 2006. Affective and pleasurable design. In: Salvendy, G. (Ed.), Handbook of Human Factors and Ergonomics, third ed. Wiley Interscience, New York, pp. 543–572.

Jordan, P.W., 1998a. An Introduction to Usability. Taylor & Francis, London.

Jordan, P.W., 1998b. Human factors for pleasure in product use. Applied Ergonomics 29 (1), 25–33.

Jordan, P.W., 2000. Designing Pleasurable Products. Taylor & Francis, London.

Khalid, H.M., 2006. Embracing diversity in user needs for affective design. Applied Ergonomics 37, 409–418.

Lavie, T., Tractinsky, N., 2004. Assessing dimensions of perceived visual aesthetics of web sites. International Journal of Human–Computer Studies 60 (3), 269–298.

Lewis, J.R., 2006. Usability testing. In: Salvendy, G. (Ed.), Handbook of Human Factors and Ergonomics. John Wiley, New York, pp. 1275–1316.

Liu, Y., 2003. Engineering aesthetics and aesthetic ergonomics: theoretical foundations and a dual-process research methodology. Ergonomics 46, 1273–1292.

McCurdy, M., Connors, C., Pyrzak, G., Kanefsky, B., Vera, A., 2006. Breaking the fidelity barrier: an examination of our current characterization of prototypes and an example of a mixed-fidelity success. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, April 22–27, 2006, Montréal, Canada. ACM, New York, pp. 1233–1242.

Nielsen, J., 1990. Paper versus computer implementations as mockup scenarios for heuristic evaluation. In: Proceedings of the IFIP TC13 Third International Conference on Human–Computer Interaction, August 27–31, 1990. North-Holland, Cambridge, UK. Amsterdam, pp. 315–320.

Norman, D.A., 2004a. Emotional Design: Why We Love (or Hate) Everyday Things. Basic Books, New York.

Norman, 2004b. Introduction to this special section on beauty, goodness, and usability. Human–Computer Interaction 19, 311–318.

Ollermann, F., 2001. Evaluation von Hypermedia-Anwendungen: Entwicklung und Validierung eines Instruments. Unpublished Diploma thesis. Institute of Work and Organisational Psychology, University of Osnabrück, Germany.

Reicherts, M., Salamin, V., Maggiori, C., Pauls, K., 2005. Psychometric characteristics of a computer-based monitoring system for emotion processing and affective states. In: Proceedings of the 10th Spanish Conference on Biometrics, 25th–27th May 2005, Oviedo.

Säde, S., Niemenen, M., Riihiaho, S., 1998. Testing usability with 3D paper prototypes – case Halton system. Applied Ergonomics 29, 67–73.

Sauer, J., Franke, H., Rüttinger, B., 2008. Designing interactive consumer products: utility of low-fidelity prototypes and effectiveness of enhanced control labelling. Applied Ergonomics 39, 71–85.

Sauer, J., Seibel, K. and Rüttinger, B. The influence of user expertise and prototype fidelity in usability tests. Manuscript submitted for publication.

Schleicher, D.J., Watt, J.D., Greguras, G.J., 2004. Reexamining the job satisfaction–performance relationship: the complexity of attitudes. Journal of Applied Psychology 89 (1), 165–177.

Sefelin, R., Tscheligi, M., Giller, V., 2003. Paper prototyping – what is it good for? A comparison of paper- and computer-based prototyping. In: Proceedings of CHI, April 5–10, 2003, Florida, USA. ACM, New York, pp. 778–779.

Shiv, B., Fedorikhin, A., 1999. Heart and mind in conflict: the interplay of affect and cognition in consumer decision making. Journal of Consumer Research 26, 278–292.

Snyder, C., 2003. Paper Prototyping: the Fast and Easy Way to Design and Refine User Interfaces. Morgan Kaufmann, San Francisco.

Tractinsky, N., 1997. Aesthetics and apparent usability: empirically assessing cultural and methodological issues. In: Proceedings of the CHI, March 22–27, 1997, Atlanta, USA. ACM, New York, pp. 115–122.

Virzi, R.A., Sokolov, J.L., Karis, D., 1996. Usability problem identification using both low- and high-fidelity prototypes. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Common Ground, April 13–18, 1996, Vancouver, Canada. ACM, New York, pp. 236–243.

Wiklund, M., Thurrot, C., Dumas, J., 1992. Does the fidelity of software prototypes affect the perception of usability?. In: Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting, October 12–16, 1992, Atlanta, USA. HFES, Santa Monica, pp. 399–403.

Willumeit, H., Gediga, G., Hamborg, K.C., 1995. Validation of the Isometrics Usability Inventory. Unpublished Research Report, Institute of Work and Organisational Psychology, University of Osnabrück, Germany.

Walker, M., Takayama, L., Landay, J.A., 2002. High-fidelity or low-fidelity, paper or computer? Choosing attributes when testing web prototypes. In: Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting, September 29–October 4, 2002, Baltimore, USA. HFES, Santa Monica, pp. 661–665.

Yamamoto, M., Lambert, D.R., 1994. The impact of product aesthetics on the evaluation of industrial products. Journal of Product Innovation Management 11 (4), 309–324.